Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/US04/041876

International filing date: 08 December 2004 (08.12.2004)

Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/528,373
Filing date: 09 December 2003 (09.12.2003)

Date of receipt at the International Bureau: 17 January 2005 (17.01.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

1269995

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

January 06, 2005

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE.

APPLICATION NUMBER: *60/528,373*
FILING DATE: *December 09, 2003*
RELATED PCT APPLICATION NUMBER: *PCT/US04/41876*

UNITED STATES PATENT AND TRADEMARK OFFICE
DEPARTMENT OF COMMERCE

Certified By

Jon W Dudas

Under Secretary
of Commerce for Intellectual Property
and Acting Director of the
Unites States Patent and Trademark Office

PTO/SB/16 (08-03)
Approved for use through 7/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No. EV 326 466 552 US

INVENTOR(S)		
Given Name (first and middle [if any])	Family Name or Surname	Residence (city and either state or foreign country)
Jack D.	NEWMAN	Berkeley, California
Neil	RENNINGER	Albany, California
Vincent J. J.	MARTIN	Kensington, California
Jay D.	KEASLING	Berkeley, California
Keith Kinkead	REILING	San Francisco, California

TITLE OF THE INVENTION (500 characters max)

METHOD FOR THE DISCOVERY, IDENTIFICATION, ISOLATION AND PEPTIDE SEQUENCING OF TERPENE SYNTHASES

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

☒ Customer Number: **23980**

OR

☐ Firm or Individual Name	REED & EBERLE LLP				
Address	800 Menlo Avenue, Suite 210				
City	Menlo Park	State	CA	ZIP	94025
Country	USA	Telephone	(650) 330-0900	Fax	(650) 330-0980

ENCLOSED APPLICATION PARTS (check all that apply)

☒ Specification *Number of Pages* 22
☒ Drawing(s) *Number of Sheets* 6
☐ Application Data Sheet. See 37 CFR 1.76
☐ CD(s), Number _____
☒ Other (specify): Return Post Card

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

	FILING FEE AMOUNT ($)
☒ Applicant claims small entity status. See 37 CFR 1.27. ☒ A check or money order is enclosed to cover the filing fees. ☒ The Commissioner is hereby authorized to charge any deficiencies in fees or credit any overpayment to Deposit Account No. 18-0580. ☐ Payment by credit card. Form PTO-2038 is attached.	$80

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☒ No.
☐ Yes, the name of the U.S. Government agency and the Government contract number are: _____

Respectfully submitted,

Date: December 9, 2003

SIGNATURE: Shelley P. Eberle
TYPED or PRINTED NAME: Shelley P. Eberle
TELEPHONE: (650) 330-0900

REGISTRATION NO.: 31,411
Docket Number: 2000-0011PRV

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

METHOD FOR THE DISCOVERY, IDENTIFICATION, ISOLATION AND PEPTIDE SEQUENCING OF TERPENE SYNTHASES

FIELD OF THE INVENTION

This invention relates to a method for discovering new terpene synthase genes from diverse organisms. More particularly, the invention relates to a mechanism-based enzyme "tagging" method to identify terpene synthase enzymes. Even more particularly, the invention relates to a method of using high resolution tandem mass spectrometry for the peptide sequencing of the "tagged" protein.

BACKGROUND OF THE INVENTION

A majority of the drugs and fine chemicals in use today are natural products or their derivatives. The source organisms (e.g., trees, marine invertebrates) of many of these natural products are neither amenable to the large-scale cultivation necessary to produce commercially viable quantities nor to genetic manipulation for increased production or derivatization of these compounds. Therefore, the natural products must be produced semi-synthetically from analogs or synthetically using conventional chemical syntheses.

Terpenes are examples of natural products and are useful as pharmaceuticals, pesticides, flavors and fragrances, and in commercial goods. Because of their structural complexity, many terpenes are currently uneconomical or impossible to synthesize. These terpenes must be either extracted from their native sources such as sponges, corals and marine microbes, or produced semi-synthetically from more abundant precursors. Such low yields and limited availability of the natural source can restrict the commercial and clinical use of terpenes. The biosynthesis of terpenes in microbes could tap the unrealized commercial and therapeutic potential of these natural resources and yield less expensive and more widely available fine chemicals and pharmaceuticals.

Terpene synthase genes (also known as terpene cyclase genes) have been isolated from plants, fungi, yeast, and terrestrial bacteria. To clone these genes, mRNA is extracted from tissues known to contain relatively high levels of terpenes. For example, terpene biosynthesis may be limited to certain plant tissue types or may be transiently elicited by wounding or

infection. This mRNA is reverse transcribed to make cDNA libraries. Because mRNA isolated from this tissue is greatly enriched in terpene synthase specific mRNA, the chances of finding a synthase gene in the cDNA library is greatly improved. In fact, cDNA libraries from terpene-producing tissue may be directly sequenced to find random clones encoding terpene synthases at a low frequency. Current methods to identify terpene synthase genes use nucleic acid probes derived from known synthase gene sequences to screen plant cDNA libraries. Probes for Southern blotting have been generated by three basic methods: (1) use of DNA of a known terpene synthase gene to identify genes with a similar nucleic acid sequence (used to isolate amorphadiene and epi-cedrol synthase genes from an *Artemisia annua* cDNA library, as well as synthase genes from potato and *Perilla frutescens*); (2) use of partial protein sequences from purified synthase enzymes to design degenerate probes for screening libraries; and (3) use of similarity-based PCR where highly conserved regions of the synthase are used to design degenerate PCR primers and amplify a region of the gene to be used as a probe (used to isolate of cDNA's encoding amorphadiene, 8-epicedrol, *ent*-kaurene, myrcene, limonene and pinene synthases, to name a few). Expression libraries have also been screened for complementation of mutant strains that are deficient in a terpene synthase, but this is only possible if synthase function is essential. Each method to clone a terpene synthase gene relies on some subset of the following criteria: the ability to target conserved sequences with a nucleic acid probe or oligonucleotides, the ability to generate enriched cDNA libraries, the ability to purify the enzyme from tissue, or the ability to express a functional enzyme in a library host.

The first committed step in the biosynthesis of all terpenes is the cyclization of a universal isoprenoid precursor molecule by a terpene synthase. The primary building block (C_5 unit) for terpenes is isopentenyl diphosphate (IPP). IPP is synthesized via two different pathways: the mevalonate pathway and the non-mevalonate, or 1-deoxyxylulose-5-phosphate (DXP) pathway (FIG. 1). The mevalonate pathway is found primarily in eukaryotes and archaea, whereas the DXP pathway is found primarily in prokaryotes, such as *E. coli*, and plastid organelles. Prenyltransferases catalyze the sequential additions of IPP to its allylic isomer dimethylallyl diphosphate (DMAPP) to form C_{10} geranyl diphosphate (GPP), C_{15} farnesyl diphosphate (FPP), C_{20} geranylgeranyl diphosphate (GGPP), and larger isoprenyl diphosphates. Amino acid substitutions near the active site can change the product distribution of the enzyme

so that an FPP synthase can be engineered to produce either GPP or GGPP. Cyclization of GPP, FPP, or GGPP by terpene synthases forms monoterpenes, sesquiterpenes, or diterpenes, respectively. All terpene synthases share a similar reaction mechanism catalyzing an intramolecular reaction of polyprenyl diphosphates.

To increase the production of IPP and the terpenoid products derived from it, researchers have over-expressed the first enzymes in the non-mevalonate pathway from *E. coli* (Wang et al. (1999) *Biotechnol Bioeng* 62(2):235-41). This engineered DXP pathway suffers from feedback regulation and carbon flux loss to cellular precursors. Because of this limitation, a complete, functional, heterologous mevalonate pathway in *E. coli* has been constructed using five genes from *Saccharomyces cerevisiae* and two from *E. coli* (Martin et al. (2003) *Nature Biotechnology* (21(7):796-801); and Keasling et al., US Serial No. 10/411,066 filed April 9, 2003 and entitled "Biosynthesis of Amoprha-4,11-diene"). This heterologous pathway was assembled into two operons, MEVT and MBIS, with specific intergenic sequences for strong gene expression. The MEVT operon contains the genes for the first three enzymes of the mevalonate pathway (*atoB* from *E. coli*, *hmgS* and *hmgR1* from *S. cerevisiae*) to convert acetyl-CoA to mevalonate. To avoid allosteric regulation, only the catalytic C-terminus of HMG-CoA reductase was used (Polakowski et al. (1998) *Appl. Microbial. Biotechnol.* 49(1):66-71). The MBIS operon encodes the last three enzymes of the mevalonate pathway (*mk, pmk, mpd* from *S. cerevisiae*) and IPP isomerase to convert mevalonate to IPP and DMAPP. FPP synthase from *E. coli* (*ispA*) was also incorporated into the mevalonate pathway operon (pMBIS) to provide an excess of substrate for sesquiterpene synthases.

The full heterologous pathway complements an *E. coli* isoprenoid auxotroph (Δ*ispC*) and allows for the high-level production of terpenes in *E. coli* when co-expressed with a terpene synthase. In addition, by replacing the FPP synthase gene (encoded in the MBIS operon) with a GPP or a GGPP synthase, a host organism has been created that is capable of over-producing mono and diterpenes, and carotenoids (*manuscript in preparation*).

Preliminary work has demonstrated the production of sesquiterpenes. The amorphadiene synthase gene (ADS) was synthesized using the *E. coli* codon preferences (Calcgene program, Hale et al. (1998) *Protein Expr Purif* 12(2):185-8). When the synthetic gene was co-expressed with the partial mevalonate pathway (MBIS) and grown in 20 mM mevalonate, *E. coli* produced

greater than 1.7 mg/L amorphadiene. When the full mevalonate pathway was added, *E. coli* produced an estimated 120 mg/L using shake flasks and short 12 hour incubation times. This production represents a >10,000-fold increase over the native plant gene using endogenous FPP. Using this strategy, mono, sesqui, and diterpenes have been produced using native plant genes and synthetic genes in *E. coli* (FIG. 2; Martin et al. (2001) *Biotechnol Bioeng* 75(5):497-503).

Diterpene synthases catalyze the conversion of GGPP to cyclic terpenes. pTrc99A-derived plasmids containing the diterpene synthases casbene (pTrcCas) or ent-kaurene (pTrcKau) synthase were co-transformed with a plasmid expressing a GGPP synthase and the resulting E. coli strains were assayed for diterpene production. The identity of the diterpenes produced was confirmed by gas chromatograph-mass spectrometry (GC-MS) analysis. The product of ent-kaurene synthase exhibited a 94% match factor with a published mass spectrum. Lacking a published spectrum for casbene, conformation of casbene production was provided by agreement of the relative abundances the major ions in the experimental fragmentation pattern to that of the published values (FIG. 3). These abundances for the bacterially-produced casbene exhibited a 97% correlation coefficient with published value for the 121, 93, 107, 136, and 272 ions (Guilford et al., (1982) *J. Am. Chem. Soc.* 104:3506-3508). When coupled to the MevT and MBIS pathways, casbene synthase produces roughly 2.5 mg/L of shake flask culture. However, even though these production levels are reasonable, higher production levels (>100 mg/L) have been observed for sesquiterpenes harvested from E. coli strains that over-produce FPP. See for example, previous work done on a bacterial strain that produces large amounts of terpenes by expressing terpene synthases in an *E. coli* strain and is engineered to provide large quantities of terpene precursors. US Patent Publication No. 20030148479 to Keasling et al. It is believed that through optimizing GGPP synthase expression and fermentation conditions, titers of diterpenes greater than 1 g/L are attainable.

The development of therapeutic terpenes is of particular interest for cancer treatment. Taxol and its derivative Taxotere are two powerful anti-cancer diterpenes used to battle not only breast and lung cancers but also the AIDS malady, Kaposi's sarcoma. Taxoid compounds are anti-mitotic agents that drive aberrant hyperpolymerization of actin and stabilization of microtubules. Microtubules are key elements of the mitotic superstructure that partitions DNA in the course of cell division. Thus, these compounds target actively growing and dividing cells

such as cancerous cells. The success of the diterpene Taxol, which was isolated from the bark of the pacific yew tree, has validated the importance of terpene natural products as chemotherapeutics. In the search for new drug leads, the vast biodiversity of the world's oceans provides a rich and diverse source for novel classes of chemicals.

Eleutherobin and sarcodictyins (FIG. 4) are potential anti-cancer compounds that share a eunicellane backbone structure and exhibit Taxoid-like modes of action. Eleutherobin was first isolated in 1995 from a soft coral (*Eleutherobia sp. Alcyonacea Alcyoniidae*), while the sarcodictyins were first isolated in 1987 from the Mediterranean stoloniferan coral *Sarcodictyon roseum*. *In vitro* assays have shown that eleutherobin and sarcodictyin A and B compete with taxol for binding to microtubules. In a series of cancer cell lines, eleutherobin and sarcodictyin A and B were shown to have IC_{50}s of 10-40, 200-400 and 200-400 nM, respectively. For comparison, Taxol exhibited low nM IC_{50}s in identical tests. Sarcodictyin A and B have been shown to be effective against taxol-resistant cancer cell lines over-expressing p-glycoprotein. Eleutherobin has been shown to have no susceptibility to several cancer cell lines with mutant, taxol-resistant tubulin. Eleutherobin and Sarcodictyins have a proven mode of action and they or derivatives hold the promise of efficacy against taxol-resistant cancers.

Despite the development of total chemical syntheses, supply limitations still hamper efforts to bring eleutherobin and the sarcodictyins to the clinic. The elegant, total synthesis routes for eleutherobin (Chen et al. (1999) *J Am Chem Soc* 121:6563; Nicolaou et al. (1999) *Chem Pharm Bull (Tokyo)* 47(9):1199-213) and the sarcodictyins (Hamel, et al.(1999) *Biochemistry* 38(17):5490-8) are far too costly to satisfy the needs of clinical trials. However, these synthesis studies have demonstrated that eleutherobin and its precursors can be used as starting materials for the chemical synthesis of derivatives (Britton et al. (2001) *J Am Chem Soc* 123(35):8632-3). Economical production of eleutherobin and the sarcodictyins or of a common structural component for use as a chemical synthon is needed to further develop these promising anticancer compounds. As an alternate source of supply, eleutherobin can be isolated from the aquarium coral *Erythropodium caribaeorum;* however, based upon the large amounts that would be required each year to meet market demand, the slow growth rates of soft coral make harvesting eleutherobin from its natural source impractical. Expressing the eleutherobin biosynthetic genes in a recombinant microorganism represents an attractive alternative for drug

production, and this strategy is currently being pursued for the anti-cancer compound bryostatin 1 (Davidson et al. (2001) *Appl Environ Microbiol* 67(10):4531-7) and the anti-molluscal agent, barbamide (Chang et al. (2002) *Gene* 296(1-2):235).

Every year numerous terpene-derived compounds with promising therapeutic properties are discovered and isolated from corals, sponges, microbes, and plants. The commercial development of these molecules can be limited by the trace quantities present in the natural sources. Therefore, there is a continuing need to develop methods of expressing the terpene biosynthetic genes in microbes, to enable scarce terpenes to be produced in the quantities required for clinical use. In spite of the progress in this field, most commercially relevant terpene synthases have not been cloned and the number of cloned terpene synthases falls far short of the number of identified terpenoid compounds. In addition, the lack of sequence identity among terpene synthases from different organisms and the low-throughput nature of current cloning methods preclude rapid screening, identification and expression of these genes. Furthermore, existing gene discovery methods are time and labor intensive and not amenable to the high-throughput cloning of terpene synthases or the generation of large gene libraries for combinatorial biosynthesis.

The present invention addresses those needs by describing a tool with which one can acquire critical genes necessary to develop of a bacterial strain capable of generating copious amounts of the desired terpene. For example, eleutherobin could potentially be produced in microbial fermentations by first isolating the genes encoding the biosynthetic pathway for the terpene chemotherapeutic.

SUMMARY OF THE INVENTION

One aspect of the invention relates to a method for identifying a terpene synthase in an environmental or other sample comprising: tagging terpene synthases present in the sample with a mechanism-based suicide substrate; identifying the tagged synthases through a tag mass shift signature using mass spectrometry; and reconstructing the synthase amino acid sequences from constituent peptides sequenced by tandem-mass spectrometry or by N-terminal sequencing of the peptides or the synthase.

Another aspect of the invention pertains to a method for purifying a terpene synthase

from a crude extract.

Yet another aspect of the invention relates to a method for the *de novo* peptide sequencing of a terpene synthase.

Yet another aspect of the invention relates to a method for obtaining the gene sequence of a terpene synthase.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates the IPP biosynthetic pathways, with the mevalonate-dependent pathway shown on the right and the non-mevalonate (DXP) pathway shown on the left. The production of FPP or GGPP can be accomplished by a single enzyme adding multiple IPPs to DMAPP or by multiple enzymes, each one catalyzing one of the steps.

FIG. 2 shows structures of mono, sesqui, and diterpenes produced in *E. coli*.

FIG. 3 are mass spectra from diterpenes extracted from engineered *E. coli*.

FIG. 4 is the structure and SAR of eleutherobin and derivatives.

FIG. 5 provides a hypothetical mechanism for the alkylation of terpene synthases by cyclopropylidene substrate analogues. The example is given for the alkylation trichodiene synthase. Adapted from Croteau et al. (1993) Archives of Biochemistry and Biophysics 307(2):397-404.

FIG. 6 is a flow-chart outlining the procedures of Examples 1-4.

FIG. 7 illustrates a proposed mechanism for the cyclization of GGPP to the eunicellane carbon backbone.

FIG. 8 illustrates the proposed mechanism for covalent modification of the eunicellane diterpene synthase with CP-GGPP.

DETAILED DESCRIPTION OF THE INVENTION

The following description of the preferred embodiments and examples are provided by way of explanation and illustration. As such, they are not to be viewed as limiting the scope of the invention as defined by the claims. Additionally, when examples are given, they are intended to be exemplary only and not to be restrictive. For example, when an example is said to "include" a specific feature, that is intended to imply that it may have that feature but not that such

examples are limited to those that include that feature. It must also be noted that, as used in this specification and the appended claims, the singular forms "a", "an" and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "an enzyme" includes a mixture of two or more such agents, reference to "a tag" includes combinations of two or more such tags, and the like.

The present invention is a method for discovering and isolating new terpene synthase genes from diverse organisms and environmental samples. A mechanism-based enzyme "tagging" method is first used to identify terpene synthase enzymes, and tandem mass spectrometry (MS) is then used for the peptide sequencing of the "tagged" protein. Thus, the isolation and peptide sequencing of numerous sesquiterpene synthases using the farnesyl diphosphate analog and a mechanistic inhibitor such as 10-cyclopropylidine farnesyl diphosphate (CP-FPP).

Briefly, to identify new terpene synthases from an organism that is known to produce terpenes (e.g., sponges, coral, plants), synthases present in crude tissue extracts are specifically alkylated or chemically modified, i.e., "tagged", using a suitable suicide substrate. For example, cyclopropyl geranyl diphosphate (GDP), farnesyl diphosphate (FDP), and geranylgeranyl diphosphate (GGDP) analogs are available for use as mechanism-based inhibitors of terpene synthases. These inhibitors covalently modify terpene synthases by alkylating amino acid residues of the synthases.

The tagged enzyme is then enriched and then amino acid sequences of portions of the enzyme are determined by tandem MS or N-terminal sequencing. The gene sequence encoding the peptides are then degeneratively reconstructed from the peptide sequence(s). Full gene sequences can then be found using the degenerate gene sequences as probes against a cDNA or genomic DNA library. Since this gene discovery method is based upon enzymatic function rather than on sequence similarity, the method of the invention has the capability to identify a broader range of terpene synthases than is possible with current, homology-based methods by capitalizing on the fact that all synthases, though dissimilar at the sequence level, perform similar chemistries on specific substrates. The sequenced synthases can then be expressed in a terpene precursor over-producing strain to produce large quantities of biosynthetically-produced, high-value terpenoids.

Accordingly, one embodiment of the invention is a method for identifying a terpene synthase in a crude extract of whole cell material comprising: tagging terpene synthases present in the extract with a mechanism-based suicide substrate; separating and identifying the tagged synthase(s) through a tag mass shift signature using nano-liquid chromatography and mass spectrometry; and reconstructing the synthase sequences from constituent peptides sequenced by tandem-mass spectrometry.

As noted above, terpene synthase genes have been identified by use of known DNA, by the use of partial protein sequences from purified synthase enzymes and by the use of similarity-based PCR. The method of the invention provides for a unique combination of these techniques used in conjunction with state of the art tandem MS to determine the sequence synthase genes and clone these genes. Using this method, synthase genes can be cloned from a marine coral, which has not been achievable by current methodologies. To date, researchers have not been able to isolate synthase genes from animal sources using DNA sequences from plants and similarity-based methodologies. This is likely due to sequence dissimilarity between plant and animal synthases, which will not present an obstacle for the MS-based method.

Mechanistic-based terpene synthase inhibitors

Terpene synthases form a highly versatile group of enzymes responsible for the biosynthesis of large families of terpene olefins and alcohols from simple polyprenyl diphosphate precursors. The enzymatic synthesis of mono, sesqui, and diterpenes by a synthase is initiated by ionization of an allylic diphosphate ester. Subsequent rearrangements of the carbocation by electrophilic cyclization, methyl or hydride migration followed by elimination of a proton (for olefin), or quenching by water (for alcohols) yields the terpenes. The ability to protect the carbocation from early cyclization termination and to chaperon the precise folding of the substrate in the synthase active site determines the ultimate structure and stereochemistry of the product(s).

Mechanism-based suicide substrates have been used in an attempt to identify important terpene synthase catalytic residues (Croteau et al. (1993) Archives of Biochemistry and Biophysics 307(2):397-404; Cane et al. (1999) *Bioorganic & Medicinal Chemistry Letters* 9(8):1127-1132). GPP, FPP, and GGPP substrate analogues containing a cyclopropyl group

function as strong mono-, sesqui-, or diterpene synthase inhibitors. The inhibitor enters the active site, and begins cyclization until forming a cyclopropyl or cyclopropylcarbonyl cation (FIG. 5). This intermediate delocalizes and stabilizes the carbocation, which can then react with nearby amino acids containing nucleophilic side chains.

The inhibitor results of Croteau and Cane, *supra*, demonstrated that cyclopropylidene analogs are substrates to synthases and are capable of alkylating (tagging) the enzymes. All twelve monoterpene synthases tested were sensitive to the cyclopropylidene geranyl diphosphate (CP-GPP) inhibitor, indicating that this mechanism-based method of "tagging" terpene synthases has broad applicability. Efforts were also directed at using the inhibitors to discover synthases from crude protein extracts and to identify the modified residues. In one instance, tagging of limonene synthase by ^{3}H-labeled CP-GPP was used to identify the enzyme from a crude protein preparation of spearmint gland extract.

Protein sequencing by tandem mass spectrometry

Mass spectrometry has become the method of choice for the sequencing and identification of proteins due to its speed, sensitivity, and the quality of data generated in the analysis. Multiple proteins can be identified or sequenced per hour, often resulting in 50-90% protein sequence coverage for a single tryptic digest. Multiple digests can yield 100% coverage with femtomole quantities of sample; attomole sensitivities are obtainable with careful tuning of the instrument. These methods, developed for purified protein samples, hold promise for analyzing complex protein mixtures.

De novo protein sequencing requires at least one purification step and a proteolytic digestion that allows the sequencing of multiple short peptide fragments. Generally, proteins are separated by SDS-polyacrylamide gel electrophoresis (PAGE) or 2D gel electrophoresis. Proteins of interest are identified on the gel, eluted from the gel, reduced and alkylated to prevent the formation of mixed disulfides, and digested with a proteolytic enzyme, usually trypsin. Eluted peptides are then separated by nano-liquid chromatography (LC) and analyzed by tandem MS. Peptide ions, created through either electrospray ionization or matrix assisted laser desorption/ionization (MALDI) are analyzed by a first stage MS to give the mass to charge ratio (m/z) of the initial peptide. In tandem MS, single peptide ions identified by the first MS, are

selected and fragmented through collision with a neutral gas. The ion products of this fragmentation can be analyzed in the second stage of mass analysis. Peptide ions most commonly fragment at the amide bond, creating ions in which the charge is retained on the N-terminus (b-ions) or the C-terminus (y-ions). The peptide sequence can be deduced from the differences in the b- and y- ion series.

Software packages accompanying tandem MS systems are able to sequence peptides co-eluting from an LC unit using automated exclusion strategies. This exclusion strategy prevents the second (sequencing) stage of MS for peptides of a specific mass that have already been sequenced. Such exclusion strategies generally allow for the simultaneous sequencing of 4 peptides over the length of an eluting LC peak (30-90 seconds). Exclusion lists can also be built automatically for entire chromatographic runs by specifying elution time and peptide m/z, allowing for the sequencing of more co-eluting peptides through multiple chromatographic runs.

The methods described herein also find utility in purifying a terpene synthase from a crude extract; in the *de novo* peptide sequencing of a terpene synthase; and for obtaining the gene sequence of a terpene synthase. Details of these aspects of the invention are described in the examples.

All patents, publications, and other published documents mentioned or referred to herein are incorporated by reference in their entireties.

It is to be understood that while the invention has been described in conjunction with the preferred specific embodiments thereof, that the foregoing description as well as the examples that follow, are intended to illustrate and not limit the scope of the invention. It should be understood by those skilled in the art that various changes may be made and equivalents may be substituted without departing from the scope of the invention, and further that other aspects, advantages and modifications will be apparent to those skilled in the art to which the invention pertains.

EXAMPLES

The practice of the present invention will employ, unless otherwise indicated, conventional techniques of synthetic organic chemistry, biochemistry, molecular biology, and the like, which are within the skill of the art. Such techniques are explained fully in the literature.

See, for example, Sambrook, et al., *Molecular Cloning: A Laboratory Manual*, 2nd edition (1989); *Oligonucleotide Synthesis* (M. J. Gait, ed., 1984); *The Practice of Peptide Synthesis* (M. Bodanszky and A. Bodanszky, 2nd ed., Springer-Verlag, New York, NY, 1994); *Nucleic Acid Hybridization* (B. D. Haines & S. J. Higgins, eds., 1984); *Methods in Enzymology* (Academic Press, Inc.); Kirk-Othmer's *Encyclopedia of Chemical Technology*; and House's *Modern Synthetic Reactions*.

The following examples describe methods that provide for the cost-effective heterologous production of eleutherobin and sarcodictiyns in an *E. coli* host. The diterpene synthase responsible for the biosynthesis of eleutherobin and the sarcodictyins in *Erythropodium caribaeorum* is first isolated. The identification of this terpene synthase provides a tool for the isolation of the enzymes responsible for further modification of the eleutherobin carbon backbone.

Cyclopropyl GGPP inhibitors are then used to isolate and sequence peptides from the terpene synthase responsible for the first step in the production of eleutherobin from GGPP. These peptide sequences are then used to design degenerate PCR primers to clone full length diterpene synthase genes from the coral sample. The genes are then expressed in an *E. coli* host to identify which synthase produces eunicellane.

In general, these experiments involve (1) confirming the presence of eleutherobin in the coral *E. caribaeorum* (or its associated symbionts) and extract protein and genomic DNA from the consortium; (2) "tagging" diterpene synthases using CP-GGPP analogs and sequence tagged tryptic peptides by LC-MS/MS using the peptide modification as a reference; (3) designing degenerate PCR primers from peptide sequences to isolate full length synthase gene sequence(s) by adapter-ligated PCR or cloning from a partial genomic library; and (4) expressing these sequences in an *E. coli* strain engineered to produce GGPP, determine the products formed by each synthesized diterpene synthase, and identify the synthase responsible for producing the eunicellane carbon skeleton found in eleutherobin. A flow-chart of these procedures is outlined in FIG. 6.

EXAMPLE 1

Isolation of Eleutherobin, nucleic acids and protein from cultured Erythropodium caribaeorum

Approximately 2 kg of cultured *Erythropodium caribaeorum* is obtained from a commercial source, such as Ocean Dreams Inc. in Tampa, Florida. This is shipped in chilled seawater to retain the coral viability. This sample is divided and used for three purposes: 1) to verify the presence of eleutherobin within the sample, 2) to obtain genomic DNA and mRNA to be used in hybridization and PCR-based identification of terpene synthases, and 3) to obtain cell lysates to be used in the functionally-based covalent modification of diterpene synthases.

Eleutherobin extraction

Eleutherobin is extracted from approximately 500 g of coral through a methanol extraction, as described in Taglialatela-Scafati et al. (2002) *Org Lett* 4(23):4085-8. Briefly, extracts are vacuum concentrated and hydrophobic organic compounds are back extracted with 50% v/w ethyl acetate. The organic layer is partitioned between hexane and 90% methanol in water, and the aqueous phase is collected. Eleutherobins are purified from the aqueous phase via elution from a normal-phase flash chromatography column at 6:4 n-hexane/ethyl acetate. The presence of eleutherobins is confirmed by UV absorbance at 290 nm (with log ε of approximately 4.0) and by liquid chromatography-mass spectrometry (LC-MS) analysis. These eleutherobins are separated using normal-phase high performance liquid chromatography (HPLC) on a silica column, eluting from CH_2Cl_2 to 30% MeOH in CH_2Cl_2, and identified using MS to confirm the expected molecular weights of eleutherobin ([M+H]+ at 656.3), desmethyleleutherobin (643.3), desacetyleleuterobin (615.3), and isoeleutherobin and z-eleutherobin (657.3).

Genomic DNA and total RNA preparation

Genomic DNA and total RNA is prepared from the cultured *E. caribaeorum* and any associated symbionts. Total RNA from *E. caribaeorum* is prepared using a method designed for "difficult sources" such as the coral *Plexaura homomalla*, bark of yew tree and marine algae (Brash et al. (1996) *J Biol Chem* 271(34):20949-57). Genomic DNA is prepared using standard extraction techniques for soft marine tissue samples, as described in (Vibede et al. (1998) *Biochem Biophys Res Commun* 252(2):497-501).

Crude enzyme preparation

The acetone powder of the coral is prepared for use in enzyme assays. Approximately 100 g of coral is homogenized in a blender for 1 minute in cold (-20°C) acetone. This mixture is centrifuged at 3000 x g for 5 minutes at 4°C and the residual solids are washed three times with cold acetone, discarding the supernatants. Protein is separated from any skeletal elements by swirling in cold acetone and decanting. The fine solids (expected to be approximately 10 g) are filtered and dried in a stream of argon and stored at -80°C. For use in enzyme assays, protein is solubilized from the powder by adding 5 mg per ml of 50 mM Tris-HCl pH 7.4.

EXAMPLE 2

Functionally-based covalent modifications of diterpene synthases in *Erythropodium caribaeorum* and its symbionts

In order to obtain probes specific for diterpene synthases in the coral sample, the terpene synthase inhibitors developed by Cane et al., *supra*, (specifically, CP-GGPP) are used to covalently modify all enzymes that cyclize GGPP. The inhibitor-tagged synthases are identified and the peptides that comprise the enzymes are sequenced using liquid chromatography-tandem mass spectrometry (LC tandem-MS) analysis, on Applied Biosystems QTRAP and QSTAR mass spectrometers.

Upon inspection of the eunicellane skeleton, the reaction mechanism is expected to be analogous to that of epi-cubenol and cadinene synthases (sesquiterpene synthases) (Benedict et al. (2001) *Plant Physiol* 125(4):1754-65). An initial ring closure at C1 to C14 would be followed by a migration of the carbocation from the C-15 position to the C-1 position due to a 1,3-hydride shift (FIG. 7). While the use of CP-GGPP was unsuccessful in covalently modifying and inhibiting taxadiene synthase (Williams et al. (2000) *Archives of Biochemistry and Biophysics* 379(1):137-146), the proposed mechanism for the eunicellane synthase differs significantly from that of taxadiene synthase. The presence of the cyclopropyl group in the inhibitor is expected to delocalize the carbocation from the C-15 position to the favored allylic C-16 and C-17 positions, making the enzyme unlikely to perform the 1,3 hydride shift, and thus unable to complete cyclization (FIG. 8). The carbocation retained on C16 or C17 would then be

free to alkylate the synthase. In the case of taxadiene synthase, the mechanism of the enzyme is likely conserved when acting upon the CP-GGPP substrate, completing cyclization such that it does not modify the enzyme (Hale et al. (1998) *Protein Expr Purif* 12(2):185-8).

Conditions for the functionally-based covalent modification reactions

Initial studies are performed using a purified diterpene synthase to determine the effects of the modification on the mass spectrometry of tagged peptides. The casbene and kaurene synthase genes have previously been cloned, and each enzyme is purified through expression as a His-tag fusion. An aliquot of each pure synthase is mixed in assay buffer with CP-GGPP. Each synthase is incubated at 30°C for up to 12 hours with an excess amount of inhibitor to ensure complete inactivation and tagging of the enzyme after the 12 hour period. Proteins from these preparations are separated and analyzed according to the protocols detailed below to identify the covalently modified peptide and amino acid.

As a further control, purified diterpene synthase is added at differing concentrations to crude cell lysates of *E. coli* and the sample is exposed to the cyclopropyl inhibitor under similar reaction conditions. Using this labeled system, conditions for enrichment of tagged synthases are determined as detailed below.

Enrichment of protein samples containing CP-GGPP tagged diterpene synthases

Enrichment simplifies MS interpretation and aids in protein sequencing. Tagged synthases are partially purified or enriched using PAGE and HPLC techniques to separate tagged from non-tagged enzymes. Radio-labeled (tritiated) inhibitor is used to follow the separation of tagged synthases from crude cell lysate.

Gel separation of tagged synthase

Crude cell lysate of an *E. coli* strain over-expressing a diterpene synthase is prepared by standard methods and exposed to 100 μM of the tritiated inhibitor. The crude enzyme preparation is separated by 1-D isoelectric focusing or 2D-gel electrophoresis, and the tagged enzyme(s) are identified as the radiolabeled spots using a Typhoon (Molecular Dynamics) multi-imager. Once spots are identified, the experiment is repeated using non-radiolabeled CP-GGPP to tag the synthases. Corresponding spots are excised, proteolyzed and extracted from the gel

(Nakayama et al. (1996) *Journal of Chromatography A* 730(1-2):279-287). Resulting peptides are analyzed by tandem-MS.

Enrichment of tagged proteins by flow-though radio-HPLC

As an alternative to enrichment by gel electrophoresis, a protocol is developed to identify and isolate the radiolabeled synthases using a flow-through radioisotope detector attached to a HPLC system. Proteins in samples of crude lysates are exposed to tritiated inhibitor and subsequently separated using a reverse-phase C4 column and eluted using a 40-90% acetonitrile gradient over 45 minutes. Eluting radiolabeled fractions are identified using a radiolabel-detector. Corresponding fractions from non-radiolabeled CP-GGPP exposed lysates are isolated and proteolytically digested. The resulting peptide fragments are combined, separated by LC, and sequenced on a high-resolution tandem MS.

Tandem-MS sequencing of tagged synthase peptide

Tagged peptides derived from terpene synthases after trypsin proteolysis are identified by detecting the presence of the tag itself. Samples are modified with the inhibitor and samples from the enzymatic reaction are taken at intervals (15 min, 30 min, 1 h, 2 h, 4 h, and 12 h) and subsequently trypsin digested. Reverse-phase liquid chromatography is used to separate peptides on a C18 micro-column (300 μm ID, 15 cm length). Ten pmoles of the proteolytic digest is loaded onto the column and eluted with a 2 to 50% acetonitrile gradient over 50 minutes.

Each peptide is fully sequenced using tandem MS. Tagged peptides are identified through the detection of sequences with "unnatural" amino acids corresponding to residues modified by the presence of the CP-GGPP tag moiety (271.2 amu). Non-natural amino acids are identified by the software in a process very similar to the identification of alkylated cysteine residues, a common practice in the analysis of tryptic peptides in proteomics. Similar techniques have been used to identify the catalytic residues for other mechanism-based inhibitors (Garcia-Alles et al. (2002) *J Biol Chem* 277(9):6934-42; Yang et al. (2000) *J Biol Chem* 275(35):26674-82).

EXAMPLE 3

Isolation of full length diterpene synthase gene sequence(s)

Full length diterpene synthases genes are isolated by PCR of adaptor-ligated cDNA or from a partial genomic library. PCR amplification of adaptor-ligated cDNA uses one gene specific primer and one primer which anneals to the double stranded adaptor ligated to the cDNA (Chenchik et al. (1996) *Biotechniques* 21(3):526-34). For purposes of this experiment, synthase specific, degenerate primers are designed to hybridize to all possible codon sets corresponding to the specific peptide sequences identified from the diterpene synthase inhibitor tagging LC-tandem-MS experiments previously described. cDNA is synthesized according to well established protocols using both random hexamer primers (for archeal and bacterial mRNA) and an oligo-dT approach (for eukaryotic mRNA) in order to assure coverage of mRNA from *Erythropodium* and its symbionts. The Marathon cDNA amplification kit from BD Biosciences Clontech (Palo Alto, CA) is employed to amplify and clone the full length cDNAs. In order to amplify both the 5' and 3' ends of the cDNA, synthase specific degenerate oligonucleotide primers, which anneal to both the top and bottom strands, are used in conjunction with adaptor specific primers. In instances where only one end of the cDNA is successfully amplified, it is sequenced to allow for the design of an alternate primer (non-degenerate). Full length cDNAs are amplified and cloned directly from adaptor-ligated cDNA using flanking 5' and 3'-synthase-specific primers established from the DNA sequences of the 5' and 3'-ends. Alternatively, the amplified 5' and 3' ends are gel purified and spliced together by PCR using the overlap extensions created by the use complementary gene specific PCR primers.

As an alternative to isolating synthase genes by PCR directly from mRNA, synthase genes are obtained from a partial genomic library. This approach has the advantage of potentially isolating gene clusters encoding several enzymes of the eleutherobin biosynthetic pathway if the eleutherobin isolated from *E. caribaeorum* is synthesized by a prokaryotic symbiont. Actinomycetes have recently been identified as common symbionts of marine organisms (Zheng et al. (2000) *FEMS Microbiol Lett* 188(1):87-91; Webster et al. (2001) *Appl Environ Microbiol* 67(1):434-44). Since Actinomycetes are known to harbor genes clusters encoding the synthesis of complex secondary metabolites including diterpenes (Dairi et al. (2001) *J Bacteriol* 183(20):6085-94), an eleutherobin biosynthetic gene cluster may be found to include all the

genes necessary to modify the olefin backbone. Since hybridization with degenerate oligonucleotides presents problems with respect to specificity and sensitivity, gene probes to be used for Southern blot analysis are generated by PCR. PCR is performed using degenerate primer pairs (forward and reverse) designed from diterpene synthase specific peptide sequences identified by MS. Sets of four pairs of primers are used in each PCR reaction, which will allow for the sampling of all possible primer pairs in fewer PCR reactions. Genomic DNA are digested with several endonucleases, Southern blotted and probed with radiolabeled nucleic acid. To generate partial libraries containing the diterpene synthase gene(s), DNA bands that hybridize to the probe(s) are excised from the gel and ligated into a pBluescript plasmid. Individual clones from the library are screened by PCR with the identical set of primers used to generate the probe.

EXAMPLE 4

Expression of diterpene synthases in E. coli and determination of synthase product

Identified diterpene synthases are cloned into pTRC99A (Pharmacia) and expressed in an *E. coli* host expressing the full mevalonate pathway and GGPP synthase (Wang et al. (1999) *Biotechnol Bioeng* 62(2):235-41). Terpene backbones produced by the cells are extracted using ethyl acetate, and purified using TLC. Products are confirmed as diterpene synthases using MS, and analyzed by NMR in order to determine the structure of each compound produced.

Overnight cultures of *E. coli* DH10B pMEVT pMBIS with GGPP synthase and a putative diterpene synthase are inoculated with stationary phase inocula, grown for two hours, followed by the expression of the mevalonate pathway and the putative diterpene synthase with 0.5 mM IPTG. Cells are grown until stationary phase. One ml samples are centrifuged, and the pellet is suspended in 1 ml of phosphate buffered saline. Diterpenes are extracted from the sample with an equal volume of ethyl acetate and subsequently purified by silica TLC (with hexane/diethyl ether 97:3 v/v). Diterpenes are located on the plate with UV light excitation of imbedded fluorescein. Diterpene spots are scraped from the plate, eluted with hexane, and analyzed by GC-MS and NMR.

It may be the case that the gene does not express well in *E. coli* due to differences in codon preference. To address this problem, genes of poorly producing putative diterpene synthases are synthesized *de novo* with *E. coli* codon preferences. After the gene sequence is

codon optimized for expression in *E. coli* using standard software (for example, Calcgene or DNAWorks), 40-basepair oligonucleotides are used for each strand of the full-length gene. The 40-bp oligonucleotides overlap by 20-bp with the oligonucleotides in the bottom strand. All of the oligonucleotides for the two strands are mixed in a single tube and assembled in a PCR thermocycler. The full-length gene is recovered from a mixture of full-length and partial products using the outer-most primers, cloned into an expression vector, transformed into the GGPP-overproducing *E. coli* strain, and screened for function by analyzing the terpene product using GC-MS analysis.

CLAIMS

WE CLAIM:

1. A method for identifying a terpene synthase in a sample comprising: tagging terpene synthases present in the sample with a mechanism-based suicide substrate; identifying the tagged synthases or constituent peptide through a tag mass shift signature using mass spectrometry; and reconstructing the synthase amino acid sequences from constituent peptides sequenced by tandem-mass spectrometry or by N-terminal sequencing of the peptides or the synthase.

2. The method of claim 1, wherein the mechanism-based suicide substrate is selected from geranyl diphosphate, farnesyl diphosphate, and geranylgeranyl diphosphate substrate analogues containing a cyclopropyl group.

3. The method of claim 2, wherein the mechanism-based suicide substrate is 10-cyclopropylidene farnesyl diphosphate.

4. The method of claim 2, wherein the mechanism-based suicide substrate is cyclopropylidene geranyl diphosphate.

5. The method of claim 2, wherein the mechanism-based suicide substrate is cyclopropylidene geranylgeranyl diphosphate.

6. The method of claim 1, wherein the mechanism-based suicide substrate covalently modifies the synthase by alkylating or attaching to amino acid residues near the mouth of the binding cleft.

7. The method of claim 1, wherein the suicide substrate is radio-labeled and the tagged synthase is identified by radio-isotope detection.

8. The method of claim 1, wherein the mRNA encoding the synthase is identified using the constituent peptide sequences and subsequently isolated and sequenced.

9. The method of claim 1, wherein the cDNA encoding for the synthase is identified using the constituent peptide sequences and subsequently isolated and sequenced.

10. The method of claim 1, wherein the chromosomal gene, which may include introns and exons, for the synthase gene is identified using the constituent peptide sequences and subsequently isolated and sequenced.

11. The method of claim 1, wherein the amino acid(s) modified by the mechanistic inhibitor are first identified and then said amino acid(s) are then mutated to any other amino acid.

12. The method of claim 11, wherein the physical properties of the synthase are altered.

13. The method of claim 1, wherein the tagged enzyme is isolated or identified by a chemical property conferred by the tag.

14. The method of claim 1, wherein the tagged enzyme is isolated or identified by the binding of the tag to a solid substrate.

15. The method of claim 1, wherein the sample is a crude extract of biological material

METHOD FOR THE DISCOVERY, IDENTIFICATION, ISOLATION AND PEPTIDE SEQUENCING OF TERPENE SYNTHASES

ABSTRACT OF THE DISCLOSURE

Method for discovering new terpene synthase genes are described. The methods utilize a mechanism-based enzyme tagging method and high resolution tandem mass spectrometry or N-terminal sequencing for the sequencing of the tagged protein.

f:\document\2000\0011prv\app.doc

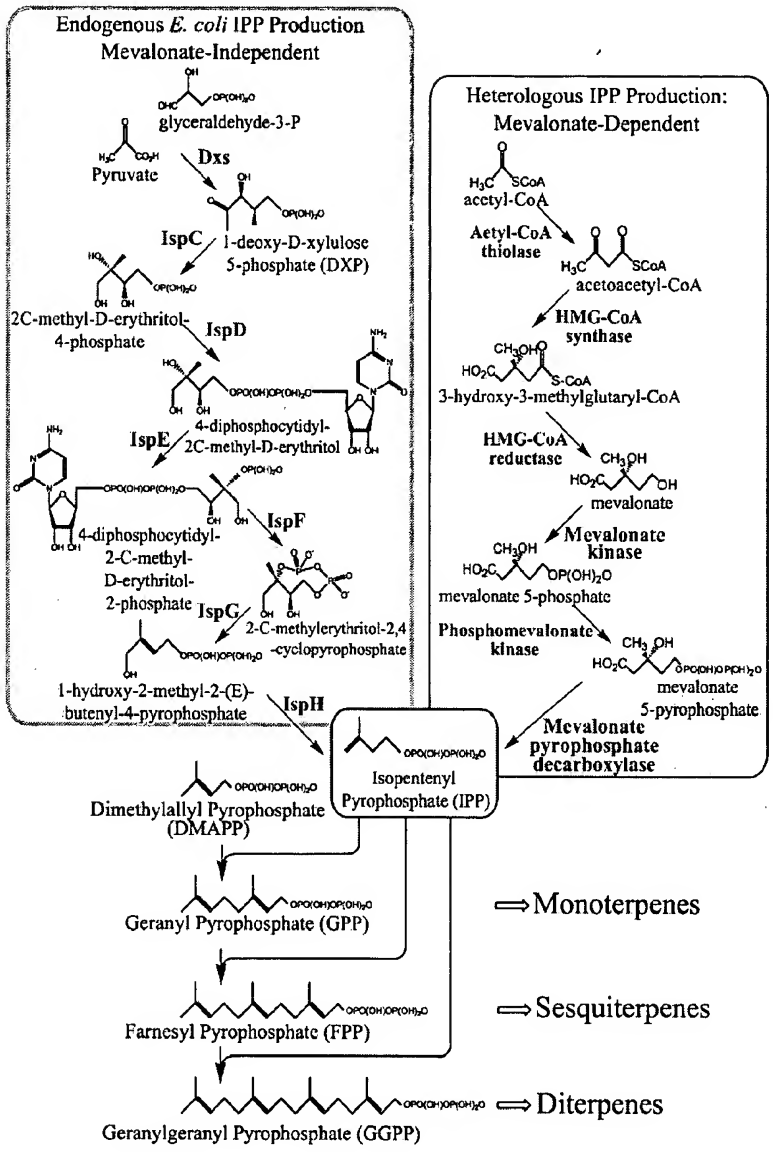

Endogenous E. coli IPP Production
Mevalonate-Independent
glyceraldehyde-3-P
Pyruvate
Dxs
1-deoxy-D-xylulose
5-phosphate (DXP)
IspC
2C-methyl-D-erythritol-
4-phosphate
IspD
4-diphosphocytidyl-
2C-methyl-D-erythritol
IspE
4-diphosphocytidyl-
2-C-methyl-
D-erythritol-
2-phosphate
IspF
2-C-methylerythritol-2,4
-cyclopyrophosphate
IspG
1-hydroxy-2-methyl-2-(E)-
butenyl-4-pyrophosphate
IspH
Heterologous IPP Production:
Mevalonate-Dependent
acetyl-CoA
Aetyl-CoA
thiolase
acetoacetyl-CoA
HMG-CoA
synthase
3-hydroxy-3-methylglutaryl-CoA
HMG-CoA
reductase
mevalonate
Mevalonate
kinase
mevalonate 5-phosphate
Phosphomevalonate
kinase
mevalonate
5-pyrophosphate
Mevalonate
pyrophosphate
decarboxylase
Isopentenyl
Pyrophosphate (IPP)
Dimethylallyl Pyrophosphate
(DMAPP)
Geranyl Pyrophosphate (GPP)
Farnesyl Pyrophosphate (FPP)
Geranylgeranyl Pyrophosphate (GGPP)
⟹Monoterpenes
⟹Sesquiterpenes
⟹Diterpenes

FIG. 1

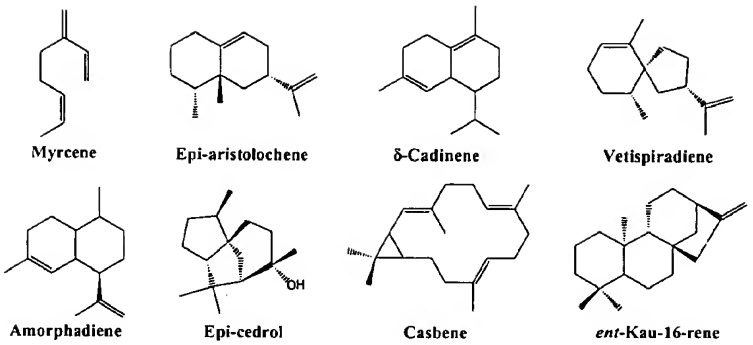

Myrcene
Epi-aristolochene
δ-Cadinene
Vetispiradiene
Amorphadiene
OH
Epi-cedrol
Casbene
ent-Kau-16-rene

FIG. 2

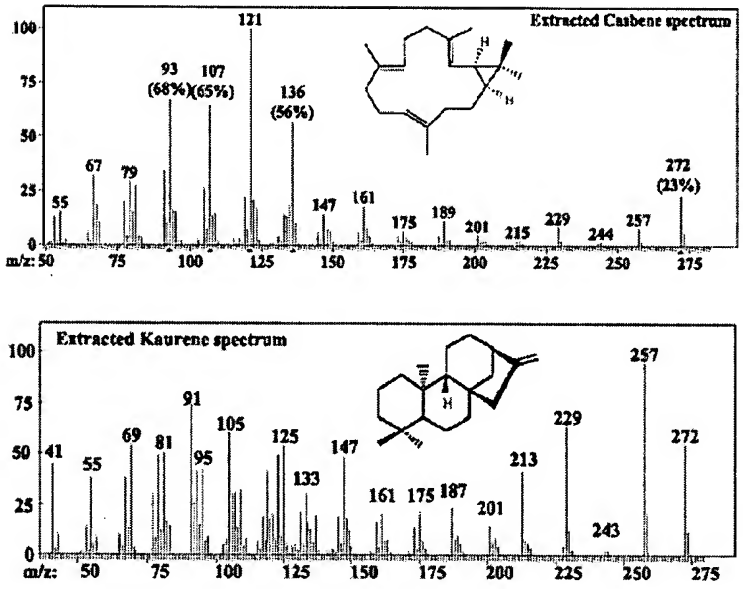
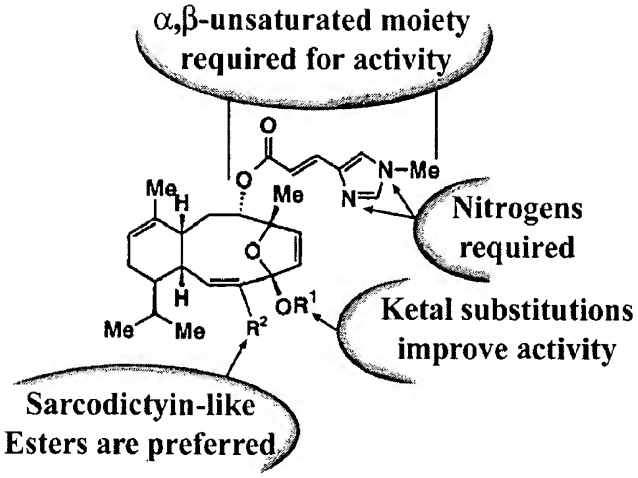
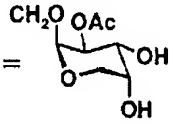

Extracted Casbene spectrum
121
93 (68%)
107 (65%)
136 (56%)
67
79
55
147
161
175
189
201
215
229
244
257
272 (23%)
m/z:
Extracted Kaurene spectrum
91
105
125
69
81
41
55
95
147
133
161
175
187
201
213
229
243
257
272
m/z:

FIG. 3



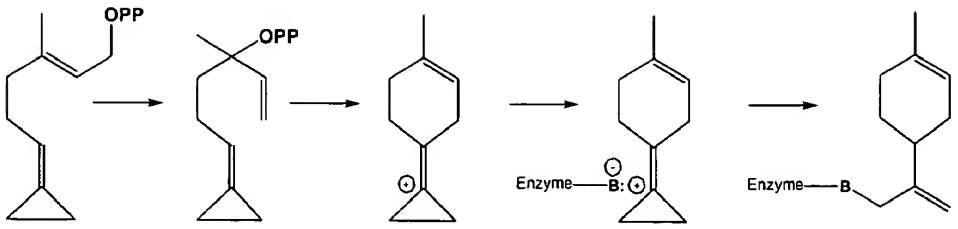
Eleutherobin: R^1 = Me, R^2 = 

Sarcodictyin A: R^1 = H, R^2 = CO_2ME

Sarcodictyin B: R^1 = H, R^2 = CO_2ET

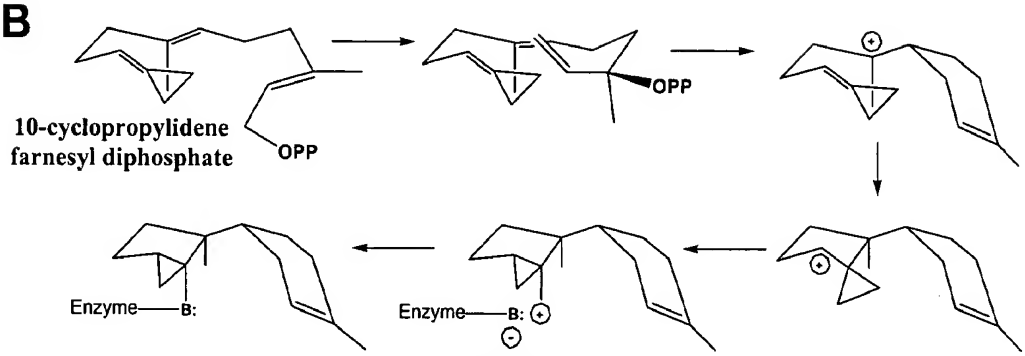
FIG. 4

A



6-cyclopropylidene
geranyl diphosphate

B



10-cyclopropylidene
farnesyl diphosphate

FIG. 5

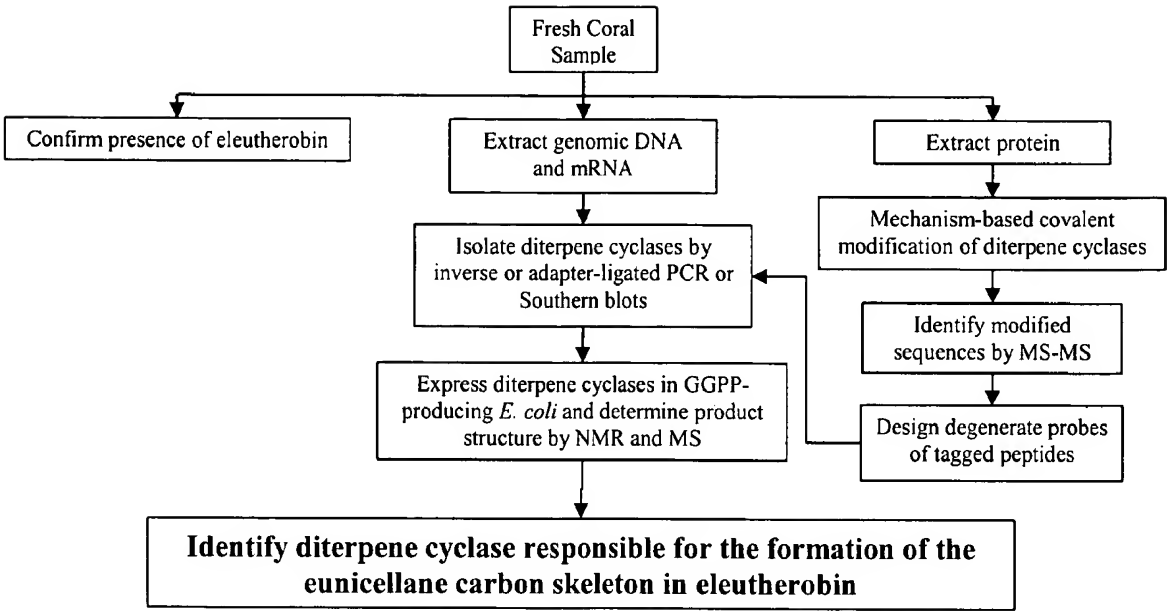

Fresh Coral Sample
Confirm presence of eleutherobin
Extract genomic DNA and mRNA
Extract protein
Mechanism-based covalent modification of diterpene cyclases
Isolate diterpene cyclases by inverse or adapter-ligated PCR or Southern blots
Identify modified sequences by MS-MS
Express diterpene cyclases in GGPP-producing E. coli and determine product structure by NMR and MS
Design degenerate probes of tagged peptides
Identify diterpene cyclase responsible for the formation of the eunicellane carbon skeleton in eleutherobin

FIG. 6

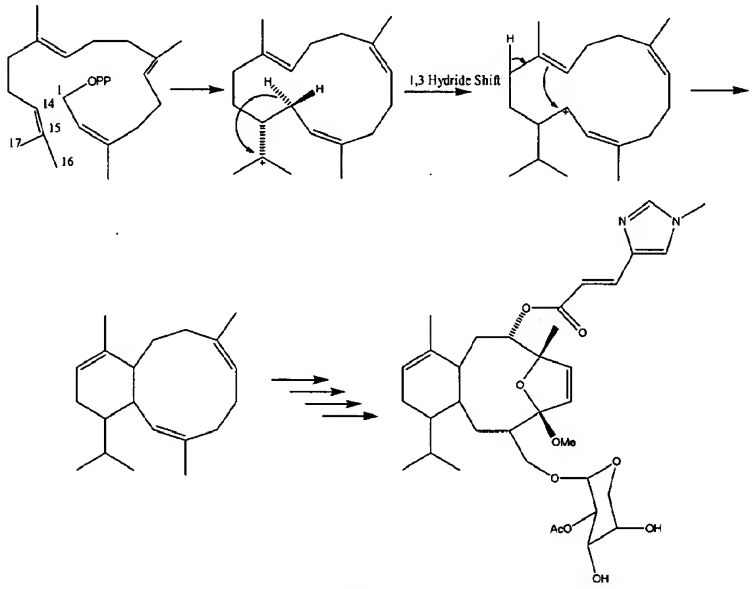

OPP
1
14
15
17
16
1,3 Hydride Shift
H
N
O
OMe
AcO
OH

FIG. 7

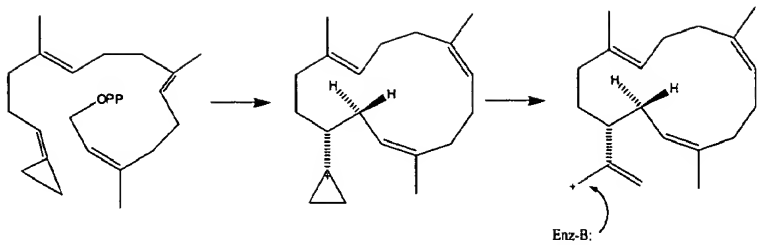


FIG. 8